ОПИСАНИЕ ГЭППИНГА В СИСТЕМЕ АВТОМАТИЧЕСКОГО ПЕРЕВОДА

Богданов А. В. (bidon@inbox.ru)

ABBYY, Москва, Россия

В работе рассматривается описание одного типа эллипсиса в системе русско-английского и англо-русского автоматического перевода. Приводится краткое определение данного типа эллипсиса (гэппинга), а также дается представление об архитектуре системы автоматического перевода. Рассматривается сам способ описания гэппинга, а затем подробно обсуждаются его плюсы и минусы. Приводятся примеры из реальных текстов, иллюстрирующие как плюсы, так и минусы описания.

Ключевые слова: эллипсис, гэппинг, машинный перевод, автоматический перевод

DESCRIPTION OF GAPPING IN A SYSTEM OF AUTOMATIC TRANSLATION

Bogdanov A. V. (bidon@inbox.ru)

ABBYY, Moscow, Russia

In this paper we discuss a description of one type of ellipsis in a Russian-English and English-Russian automatic translation system. Short definition of this type of ellipsis (gapping) is given, as well as an overview about architecture of a system of automatic translation. The method of gapping description itself is considered and then its advantages and disadvantages are discussed in detail. Examples from real texts illustrating advantages, as well as disadvantages of description are given.

Key words: ellipsis, gapping, machine translation

1. Рабочее определение гэппинга и теоретические вопросы

Гэппингом (англ. *gapping*) мы называем тип эллипсиса, при котором сокращается вершина одной или нескольких неначальных сочиненных составляющих. При этом могут подвергаться частичному сокращению и зависимые этой сокращенной вершины.

(1) *Петя любит Машу, а Маша — Петю.*

(2) *Традиционный салат «Оливье» будет стоять на столах 64% россиян, соленья — 54%, селедка под шубой — 46%.*

В (1) сокращается глагол (*любит*), являющийся вершиной правого конъюнкта сочинительной цепочки, при этом все его зависимые (*Маша, Петю*) остаются на месте. В (2) сокращается два глагола (*будут (будет) стоять*), являющиеся также вершинами неначальных конъюнктов. При этом подлежащее этих глаголов остается на месте, а другие их зависимые также подвергаются частичному эллипсису — на поверхности остаются лишь %, являющиеся прадочерними составляющими для сокращенных глаголов[1].

Сокращению не может подвергаться левый конъюнкт:

(3) ** Петя Машу, а Маша любит Петю.*

Гэппинг встречается во многих языках, и ему посвящено множество исследований, среди которых: [Ross 1970], [Jackendoff 1971], [Kuno 1976], [Neijt 1979], [Winkler 2005] и др.

Одна из самых сложных проблем при теоретическом описании гэппинга — тот факт, что сокращению может подвергаться материал, не образующий составляющую (как в (2)). Однако известны и строгие ограничения, накладывающиеся на гэппинг, — так, в [Ross 1970] впервые отмечается, что при гэппинге сокращению подлежит только такой материал, который образует непрерывную подчинительную структуру.

(4) а. *Маша хочет научиться курить сигареты, …*
б. *а Петя — научиться курить сигары.*
в. *а Петя — курить сигары.*
г. *а Петя — сигары.*
д. ** а Петя — научиться сигары.*

В (4б, в, г) сокращению подвергается непрерывная подчинительная структура, включающая правый конъюнкт — *хочет*; *хочет научиться* и *хочет*

---

[1] В нашем понимании и в соответствии с приведенным нами определением, тип эллипсиса, обычно называемый left node raising (*Chomsky sent an offprint to Harris and a manuscript to his editor*), также является частным случаем гэппинга.

научиться курить. Тогда как в (4д) сокращению подвергается правый конъюнкт и его зависимое, не образующее с ним непрерывную подчинительную структуру — *хочет <...> курить*. И в этом случае сокращение оказывается невозможным.

2. Архитектура системы и проблемы описания гэппинга

Прежде чем говорить о способах описания гэппинга, кратко опишем архитектуру системы автоматического перевода, о которой пойдет речь.

Данная система перевода состоит из двух модулей — модуля анализа и модуля синтеза. На вход модуля анализа поступает текст на естественном языке, на выходе имеем семантическую структуру (на специальном семантическом языке); на вход модуля синтеза поступает семантическая структура, на выходе имеем текст на естественном языке.

Семантическая структура, упрощенно говоря, представляет собой дерево, в узлах которого находятся семантические концепты (обобщенные семантические единицы, не являющиеся объектами естественного языка), а связи помечены семантическими ролями.

(5) а.	*Мальчик*	*дал*	*девочке*	*яблоко.*
б.	[[$_{\text{Subject}}$ мальчик]	дать	[$_{\text{Object\_Dative}}$ девочка]	[$_{\text{Object\_Direct}}$ яблоко]]
в.	[[$_{\text{Agent}}$ BOY]	TO_GIVE	[$_{\text{Possessor}}$ GIRL]	[$_{\text{Object}}$ APPLE]]
г.	[[$_{\text{Subject}}$ boy]	give	[$_{\text{Object\_Indirect\_to}}$ girl]	[$_{\text{Object\_Direct}}$ apple]]
д.	*The boy gave an apple to a girl.*			

В (5а) представлен входной текст на русском языке. В (5б) — синтаксическая структура входного текста (промежуточный результат работы модуля анализа). Все связи в синтаксической структуре помечены типами (так называемыми поверхностными позициями). В (5в) представлена семантическая структура, в которой связи помечены семантическими ролями (концепты записаны заглавными латинскими буквами). В (5г) — синтаксическая структура выходного текста (с поверхностными позициями выходного языка), наконец, в (5д) — выходной текст на английском языке.

Очевидно, что для такой системы любой тип эллипсиса представляет некоторую сложность, так как все словоформы во входном тексте должны быть интерпретированы (им должны быть приписаны поверхностные позиции и должна найтись вершина, их присоединяющая). В простых случаях эллипсиса это решается восстановлением в месте сокращения некоторой условной вершины, к которой присоединяются имеющиеся во входном тексте зависимые.

Так, например, в случае регулярного эллипсиса *быть* в русском языке можно описать условия, при которых нужно всегда восстанавливать данный глагол.

(6) а. *Мальчик* *готов.*
б. [[$_{\text{Subject}}$ мальчик] <быть> [$_{\text{Complement}}$ готов]]
в. [[$_{\text{Object}}$ BOY] TO_BE [$_{\text{State}}$ READINESS]]

В (6) показан переход от сокращенной структуры входного текста (6а) к синтаксической структуре с восстановленным *быть* (6б) и далее к семантической структуре (6в). В упрощенном виде условия для восстановления *быть* в (6) выглядят так:

(6') […[Subject] …<быть>… [Complement]…];

Модуль анализа интерпретирует шаблон, записанный в (6') так: если во входном тексте имеется фрагмент, подходящий под этот шаблон, то нужно построить синтаксическую структуру с вершиной *быть* и данными зависимыми.

Простые случаи гэппинга вроде (1) можно описывать подобным образом. Например, можно использовать для этого такой шаблон:

(1') […[Subject]…Verb…[Object_Direct]…] & […[Subject]…<Verb>… [Object_Direct]…]

При этом на месте сокращенного глагола можно восстанавливать лексему глагола, представленного в левом конъюнкте.

Однако шаблоны типа (1') неэффективны для описания гэппинга по следующим причинам:

- Зависимые сокращенного конъюнкта (как и левого конъюнкта) могут быть самых разных типов — понадобятся шаблоны для разных комбинаций поверхностных позиций или какая-либо обобщенная запись поверхностных позиций.
- В большинстве случаев гэппинга имеется параллелизм поверхностных позиций зависимых левого конъюнкта и сокращенного конъюнкта — понадобится каким-то образом в шаблоне учитывать совпадение поверхностных позиций двух конъюнктов.
- Сокращению может подвергаться более глубокая структура, как в (2), и для таких случаев в шаблонах придется указывать все сокращенные элементы, которых может быть потенциально очень много.

Приведенные причины делают практически невозможным написание обобщенного шаблона типа (1') для любых случаев гэппинга.

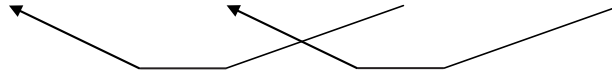
3. Описание гэппинга через контролирующие поверхностные позиции

Учитывая все вышесказанное, а также тот факт, что на гэппинг в русском и в английском языке накладываются схожие ограничения, была разработана

другая система описания гэппинга — через контролирующие поверхностные позиции.

В упрощенном виде эта система выглядит следующим образом. Имеются две (в расширенном варианте — больше двух) технические поверхностные позиции с широким заполнением — то есть, такие, что в них могут попадать составляющие разных грамматических категорий (именные группы, глагольные группы и др.) и в разной грамматической форме (в разных падежах, с разными предлогами, союзами и т. п.). В этих позициях прописано свойство контроля. Контроль устанавливается между этими позициями и дочерними или прадочерними зависимыми левого конъюнкта. Покажем это на схеме.

(7) […[dependent]…Verb…[dependent]...] & [[Remnant1] <Verb> [Remnant2]]



В (7) показан шаблон с контролирующими поверхностными позициями. Ярлыком dependent обозначено множество поверхностных позиций для зависимых левого конъюнкта (будем записывать такие множества с маленькой буквы, в отличие от поверхностных позиций). Многоточиями обозначены другие возможные зависимые. Обозначения Remnant1 и Remnant2 — это технические поверхностные позиции для гэппинга, которые и являются контролирующими. Контроль от них идет в зависимые левого конъюнкта, он условно обозначен стрелками. Условием на установление контроля может быть согласование по определенному списку грамматических категорий (например, для существительных — по падежу и предлогу).

Покажем, как работает анализ и синтез примера (8а) по шаблону (7).

(8) а. *Мальчик рассказал девочке про синтаксис, а мне — про морфологию.*

б. [[$_{\text{Subject}}$ мальчик] рассказать [$_{\text{Object\_Dative}}$ девочка; w] [$_{\text{Object\_Indirect\_Про}}$ синтаксис; x]] & [[$_{\text{Remnant1}}$ я; y] <Ellipted> [$_{\text{Remnant2}}$ морфология; z]];
y → w <падеж; предлог>
z → x <падеж; предлог>

в. [[$_{\text{Agent}}$ BOY] TO_TELL [$_{\text{Addressee}}$ GIRL; w] [$_{\text{Theme}}$ SYNTAX; x]] & [[$_{\text{Remnant}}$ I; y] <Ellipted> [$_{\text{Remnant}}$ MORPHOLOGY; z]];
y → w
z → x

г. [[$_{\text{Subject}}$ boy] tell [$_{\text{Object\_Dative}}$ girl; w] [$_{\text{Object\_Indirect\_About}}$ syntax; x]] & [[$_{\text{Remnant1}}$ I; y] <Ellipted> [$_{\text{Remnant2}}$ morphology; z]];
y → w <падеж; предлог>
z → x <падеж; предлог>

д. *The boy told the girl about syntax and me — about morphology.*

В (8б) приведена синтаксическая структура, полученная в результате применения шаблона (7) к (8а). Латинскими строчными буквами помечены

узлы структуры, участвующие в контроле. После скобочной записи перечислены две связи контроля — из поверхностной позиции Remnant1 в поверхностную позицию Object_Dative и из поверхностной позиции Remnant2 в поверхностную позицию Object_Indirect_Про. При установлении этих связей проверяется совпадение значений категорий падежа и предлога у контролера и мишени.

В (8в) приведена семантическая структура — результат работы модуля анализа. В ней сохраняются обе связи контроля, а зависимые сокращенного глагола получают технические семантические роли Remnant.

В (8г) приведена синтаксическая структура выходного языка. Обе связи контроля в ней сохраняются и реализуются через согласование по падежу и предлогу между контролером и мишенью. В результате получаем выходной текст (8д), в котором грамматическая форма *me* и *about morphology* — это результат работы согласования, то есть копирования значений соответствующих категорий с мишеней данных связей.

Такой способ описания гэппинга удобен тем, что позволяет в сложных случаях (как, например, в (2)) не восстанавливать промежуточные составляющие (между сокращенной вершиной и «остатками» гэппинга), но при этом синтезировать правильную грамматическую форму «остатков».

Продемонстрируем, как работает такое описание на более сложном примере с сокращенными промежуточными составляющими.

(9) а. *Информация синтетических счетов отражается в финансовой отчетности, а субсчетов — в приложениях к финансовой отчетности.*

б. [[$_{\text{Subject}}$информация[$_{\text{GenitivePostModifier}}$счет;w]]отражать[$_{\text{Adjunct_Locative}}$отчетность; x]] & [[$_{\text{Remnant1}}$ субсчет; y] <Ellipted> [$_{\text{Remnant2}}$ приложение; z]]
y → w <падеж; предлог>
z → x <падеж; предлог>

в. [[$_{\text{Object}}$ INFORMATION [$_{\text{Possessor\_Metaphoric}}$ BANK_ACCOUNT; w]] TO_MIRROR [$_{\text{MetaphoricLocative}}$ FINANCIAL_REPORTING; x]] & [[$_{\text{Remnant}}$ SUBACCOUNT; y] <Ellipted> [$_{\text{Remnant}}$ SUPPLEMENT; z]]
y → w
z → x

г. [[$_{\text{Subject}}$ information [$_{\text{OfPostModifier}}$ account; w]] reflect [$_{\text{Adjunct\_Locative}}$ statement; x]] & [[$_{\text{Remnant1}}$ subaccount; y] <Ellipted> [$_{\text{Remnant2}}$ annex; z]]
y → w <падеж; предлог>
z → x <падеж; предлог>

д. *Information of control accounts is reflected in financial statements and of subaccounts — in annexes to financial statements.*

В (9б–в) показана работа модулей анализа и синтеза с теми же обозначениями, что и в предыдущих примерах. Некоторые составляющие из (9а) для краткости опущены. В отличие от (8), контроль в (9б) устанавливается не в дочернюю левого конъюнкта, а в прадочернюю (*счетов*), так как выполняется условие согласования именно с этой составляющей. Далее

на синтезе в (9г) этот контроль обеспечивает согласование по падежу и предлогу между *subaccount* и *account*, а также между *annex* и *statement*. Благодаря чему получаем нужные грамматические формы *of subaccounts* и *in annexes*.

Для того, чтобы был возможен анализ типа (9б), в правилах контроля наряду с согласованием должны быть заданы возможные пути до мишени. Путем в данном случае мы называем цепочку поверхностных позиций последовательно зависящих друг от друга. Путь, обеспечивающий анализ (9б), выглядит так:

(10) Subject.GenitivePostModifier;

Путь, обеспечивающий синтез (9г), выглядит так:

(11) Subject.OfPostModifier;

Возможен гэппинг и с более глубоким сокращением, для которого могут понадобиться более длинные пути. Например:

(12) а. *Торговцам предоставили возможность заниматься* ***торговлей****, ремесленникам — своим ремеслом.*
б. Object_Direct.Clause_Infinitive_NoControl.Object_Indirect;

(13) а. *Rate of growth of credits to nonfinancial borrowers* ***in rubles*** *was of 101,9%, in foreign currency of 103,3%.*
б. Subject.OfPostModifier.OfPostModifier.Object_Indirect_in;

В (12) и (13) в пунктах а. представлены примеры из реальных текстов, а в пунктах б. — пути до мишени контроля, необходимые для описания гэппинга в этих примерах. Жирным выделены мишени контроля.

Очевидно, что в таких примерах восстановление всех сокращенных промежуточных составляющих было бы более трудоемкой и сложной задачей.

4. Возможные расширения предложенного описания

Описанную выше систему можно расширить несколькими способами.

Во-первых, число контролирующих позиций может меняться — встречаются случаи, когда параллелизм наблюдается только в одной паре поверхностных позиций, или наоборот — более чем в двух парах.

(14) *Объем привлеченных Сбербанком России депозитов увеличился* ***на 38,2%****, а в остальных кредитных организациях —* ***на 65,4%****.*

(15) *Внутренний кредит увеличился* ***на 544,6 млрд. руб.****, а за II квартал 2006 г. —* ***на 126,7 млрд. руб.***

В (14) и (15) показаны примеры, в которых только один «остаток» гэппинга параллелен одной из зависимых левого конъюнкта — *на 65,4%* и *на 126,7 млрд. руб.*. Вторая зависимая сокращенного глагола (*в остальных кредитных организациях* и *за II квартал 2006 г.*) в обоих примерах не имеет никакого коррелята в левой части. Точнее, в (14) таким коррелятом можно считать составляющую *Сбербанком России,* однако из-за отсутствия согласования по каким-либо категориям установить связь контроля между этими составляющими невозможно. Поэтому к примерам (14) и (15) не может быть применен шаблон типа (7). Однако подобные примеры можно описать с помощью шаблона с одной контролирующей позицией.

(16) […[dependent]…Verb…[dependent]...] & [[dependent] <Verb> [Remnant]]

В шаблоне (16) у сокращенного глагола (правого конъюнкта) указаны две зависимые — одна из них (dependent) может занимать любую поверхностную позицию из соответствующего списка (который можно дополнительно сузить), а вторая зависимая попадает в поверхностную позицию Remnant, из которой обязательно должен установиться контроль.

Другим возможным расширением предложенного описания может быть его применение к подчинительным конструкциям.

(17) *Молодые люди сегодня покидают дом быстрее, чем их коллеги поколение назад.*

(18) *Если в 1998 году у нас трудилось 1,1 тысяч иностранцев, то в прошлом году — 1,6.*

В (17) и (18) показаны примеры с эллипсисом, аналогичным гэппингу. Их отличие лишь в том, что сокращенный глагол не сочинен с главным, а подчинен ему. Такие случаи тоже могут быть описаны через контролирующие позиции с помощью шаблона вида:

(19) […[dependent]…Verb…[dependent]... [[Remnant1] <Verb> [Remnant]]]

После применения шаблона (19) к (17) должна построиться синтаксическая структура с двумя связями: (*их коллеги*) → (*Молодые люди*) и (*поколение назад*) → (*сегодня*). Аналогично для (18): (*в прошлом году*) → (*в 1998 году*) и (*1,6*) → (*1,1*).

Следует при этом принять во внимание тот факт, что возможные пути контроля, указанные в поверхностных позициях Remnant, в случае шаблона (7) идут от конъюнкта сокращенного глагола, а в случае шаблона (19) эти пути идут от родительской составляющей сокращенного глагола. Это различие можно указать, например, в самих шаблонах.

5. Недостатки предложенного описания

Предложенное описание гэппинга хорошо работает только в тех случаях, когда синтаксические структуры для данного примера на входном и выходном языке подобны, как во всех приведенных выше примерах. Если же синтаксическая структура входного языка при переводе должна каким-то образом трансформироваться, то предложенное описание может выдать плохой результат. Приведем примеры.

(20) *The boy was asleep.* → *Мальчик спал.*

Для перевода (20) в модуле анализа используется правило, которое производит следующее преобразование синтаксической структуры:

(21) [BE [$_{\text{Complement}}$ TO_SLEEP]] => [TO_SLEEP]

Теперь приведем пример с гэппингом, в котором *asleep* контролируется из позиции Remnant.

(22) а. *The boy was asleep and the girl — awake.*
б. [[$_{\text{Subject}}$ boy; w] be [$_{\text{Complement}}$ sleep; x]] & [[$_{\text{Remnant1}}$ girl; y] <Ellipted> [$_{\text{Remnant2}}$ awake; z]]
y → w
z → x
в. [[$_{\text{Experiencer}}$ BOY; w] TO_SLEEP; x] & [[$_{\text{Remnant}}$ GIRL; y] <Ellipted> [$_{\text{Remnant}}$ TO_STAY_AWAKE; z]]
y → w
z → x

В (22а) приведен пример, в котором сокращен правый конъюнкт, а одним из его оставшихся зависимых является *awake*, параллельное *asleep* в левом конъюнкте. В (22б) приведена синтаксическая структура, получившаяся в результате применения шаблона (7). Далее действует правило (21), заменяющее BE на TO_SLEEP и удаляющее составляющую в поверхностной позиции complement. После чего имеем семантическую структуру (22в), в которой связь контроля от TO_STAY_AWAKE идет к TO_SLEEP — то есть, к левому конъюнкту, а не к какому-либо из его зависимых. Такая семантическая структура не может синтезироваться стандартным образом, так как пути контроля в позициях Remnant не могут быть нулевыми.

Другим типичным примером недостатка предложенной системы описания гэппинга являются случаи, в которых между двумя «остатками» гэппинга должна быть какая-либо связь. Приведем пример.

(23) а. *The boy was stupid.*
б. [[$_{\text{Subject}}$ boy; x] be [$_{\text{Complement}}$ stupid; y]]
y → x

В (23) приведен пример, в котором между двумя зависимыми глагола *be* устанавливается дополнительная связь контроля. Она необходима в частности для того, чтобы согласовывать подлежащее и комплемент в таких конструкциях по роду и числу в русском языке (*Мальчик был глуп**ым***.). Этот контроль прописан в поверхностной позиции Complement, то есть его установление обязательно для этой позиции.

Если подобные конструкции встречаются в контексте гэппинга, то возникает следующая проблема.

(24) а. *The boy was stupid and the girl — clever.*
б. [[$_{\text{Subject}}$ boy; w] be [$_{\text{Complement}}$ stupid; x]] & [[$_{\text{Remnant1}}$ girl; y] <Ellipted> [$_{\text{Remnant2}}$ clever; z]]
y → w (контроль прописан в позиции Remnant1)
z → x (контроль прописан в позиции Remnant2)
x → w (контроль прописан в позиции Complement)
в. [[$_{\text{Object}}$ BOY; w] BE [$_{\text{State}}$ STUPID; x]] & [[$_{\text{Remnant}}$ GIRL; y] <Ellipted> [$_{\text{Remnant}}$ CLEVER; z]]
y → w
z → x
x → w
г. [[$_{\text{Subject}}$ мальчик; w] быть [$_{\text{Complement}}$ глупый; x]] & [[$_{\text{Remnant1}}$ девочка; y] <Ellipted> [$_{\text{Remnant2}}$ умный; z]]
y → w (контроль прописан в позиции Remnant1) <падеж; предлог>
z → x (контроль прописан в позиции Remnant2) <падеж>
x → w (контроль прописан в позиции Complement) <род; число>
д. *Мальчик был глупым, а девочка умн**ым***.

В (24б) приведена синтаксическая структура для (24а), полученная в результате применения шаблона (7). Перечислены все три связи контроля — две связи из контролирующих позиций гэппинга и третья связь — из позиции Complement в подлежащее. При переходе к семантической структуре (24в) все три связи сохраняются. Далее синтезируется синтаксическая стркутура (24г), в которой реализуется связь из позиции Remnant1 с согласованием по падежу и предлогу (в результате чего имеем форму *девочка*). Также реализуется связь из позиции Remnant2 с согласованием только по падежу (прилагательные при гэппинге не согласуются ни по каким другим категориям) и связь из позиции Complement с согласованием по роду и числу (в результате чего имеем форму *глупым*). При этом род и число для составляющей *умным* взять неоткуда. Фактически эти категории остаются неопределены, а в (24д) приведена форма с дефолтными значениями.

Другими словами, в конструкциях такого типа не хватает одной связи контроля — между *clever* и *girl*. Но ввести эту связь в структуру проблематично, так как она инициируется позицией Complement, а в данном случае *clever* подключается не в позицию Complement, а в позицию Remnant2.

Подобная проблема может возникнуть во всех случаях, в которых между двумя «остатками» гэппинга должна быть какая-либо дополнительная связь.

References

1. *Jackendoff, Ray* (1971) Gapping and related rules. Linguistic Inquiry 2, 21–35.
2. *Kuno, S.* (1976) Gapping: A functional analysis. Linguistic Inquiry 7, 300–318.
3. *Neijt, A. H.* (1979) Gapping. A contribution to sentence grammar, Foris:Dordrecht
4. *Ross, J.* (1970) Gapping and the order of constituents. In M. Bierwisch & K. Heidolph (eds.), Progress in linguistics: A collection of papers, pp. 249–259, The Hague: Mouton.
5. *Winkler, S.* (2005) Ellipsis and Focus in Generative Grammar. Mouton de Gruyter: Berlin/ New York.